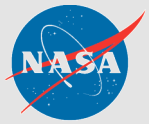


National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Mining Massive Earth Science Data Sets for Large Scale Structure

Amy Braverman and Eric Fetzer

Jet Propulsion Laboratory,

California Institute of Technology

Mail Stop 126-347

4800 Oak Grove Drive

Pasadena, CA 91109

Email: Amy.Braverman@jpl.nasa.gov

ACKNOWLEDGEMENTS

- The Atmospheric Infrared Sounder (AIRS) Project at JPL.
- The Multi-angle Imaging SpectroRadiometer Project (MISR) at JPL.
- NASA's Earth-Sun Technology Office's Advanced Information Systems Technology Program (ESTO AIST)

Outline

- **Background and Problem Statement**
- **Approach and Methodology**
- **Data Mining Example**
- **Conclusions**

Background and Problem Statement

- **NASA's Earth Observing System (EOS) collects massive, distributed, heterogeneous data "sets".**
- **EOS data sets big and complex.**
  - **Phenomena of interest at different spatial scales, but we don't know what we don't know.**
- **How can we quantitatively characterize and compare large-scale structure (i.e. Level 3) in time and space?**

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

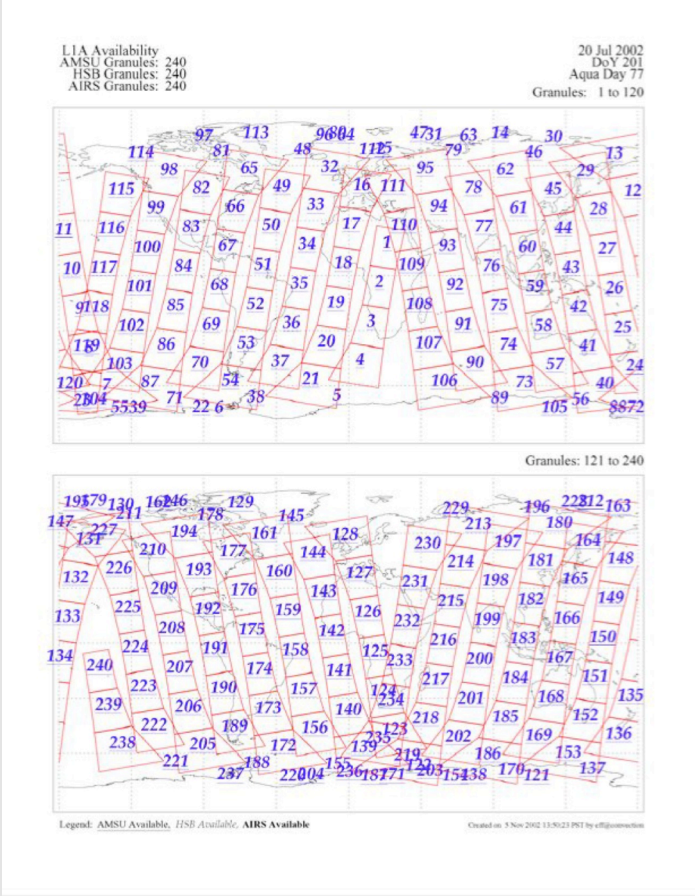
Background

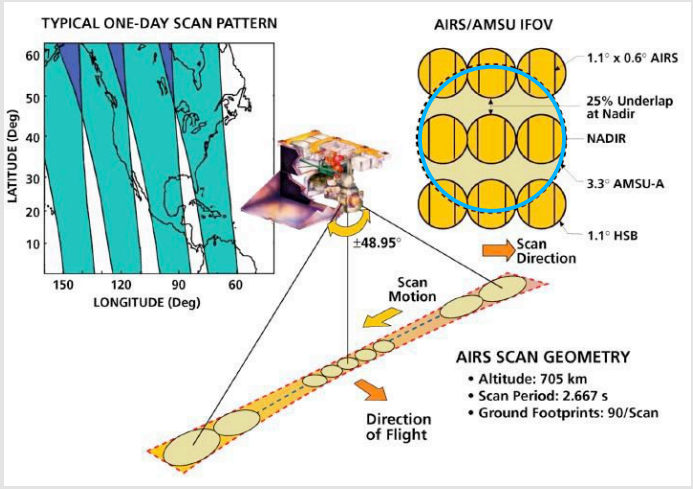
- ***Assertion: empirical probability distributions derived from the data are signatures of physical processes.***
- **Distributions defined on different space-time windows can be compared, and differences or changes attributed to physical processes.**
- **Approach:**
  - **partition data on coarse, spatio-temporal grid (monthly five-degree)**
  - **summarize the data in each grid cell by a multivariate distribution estimate, i.e the output of a clustering algorithm.**
  - **use a well-defined "distance" between probability distributions as a basis for data mining algorithms.**

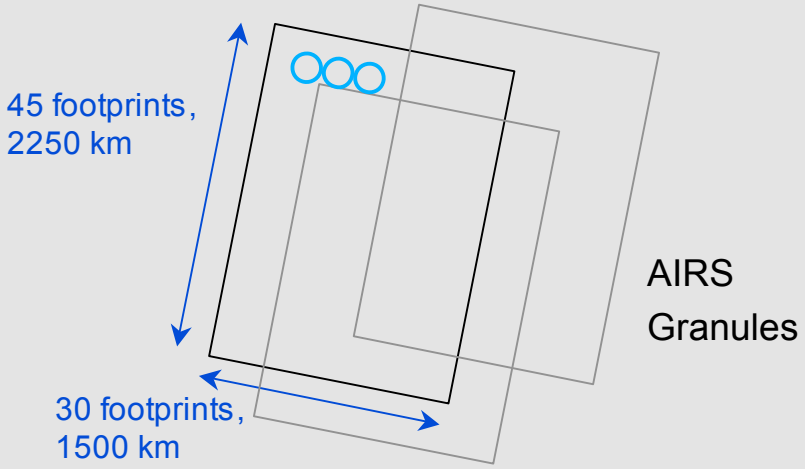
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Atmospheric Infrared Sounder (AIRS) Data







National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

AIRS Data

AIRS Variables:

Indices	*Field*
1-11:	atmospheric temperature at 11 altitudes
12-22:	atmospheric humidity at 11 altitudes
13-32:	cloud fraction at 10 altitudes (excludes lowest)
33:	land fraction
34:	granule type (ascending or descending)
35:	quality flag

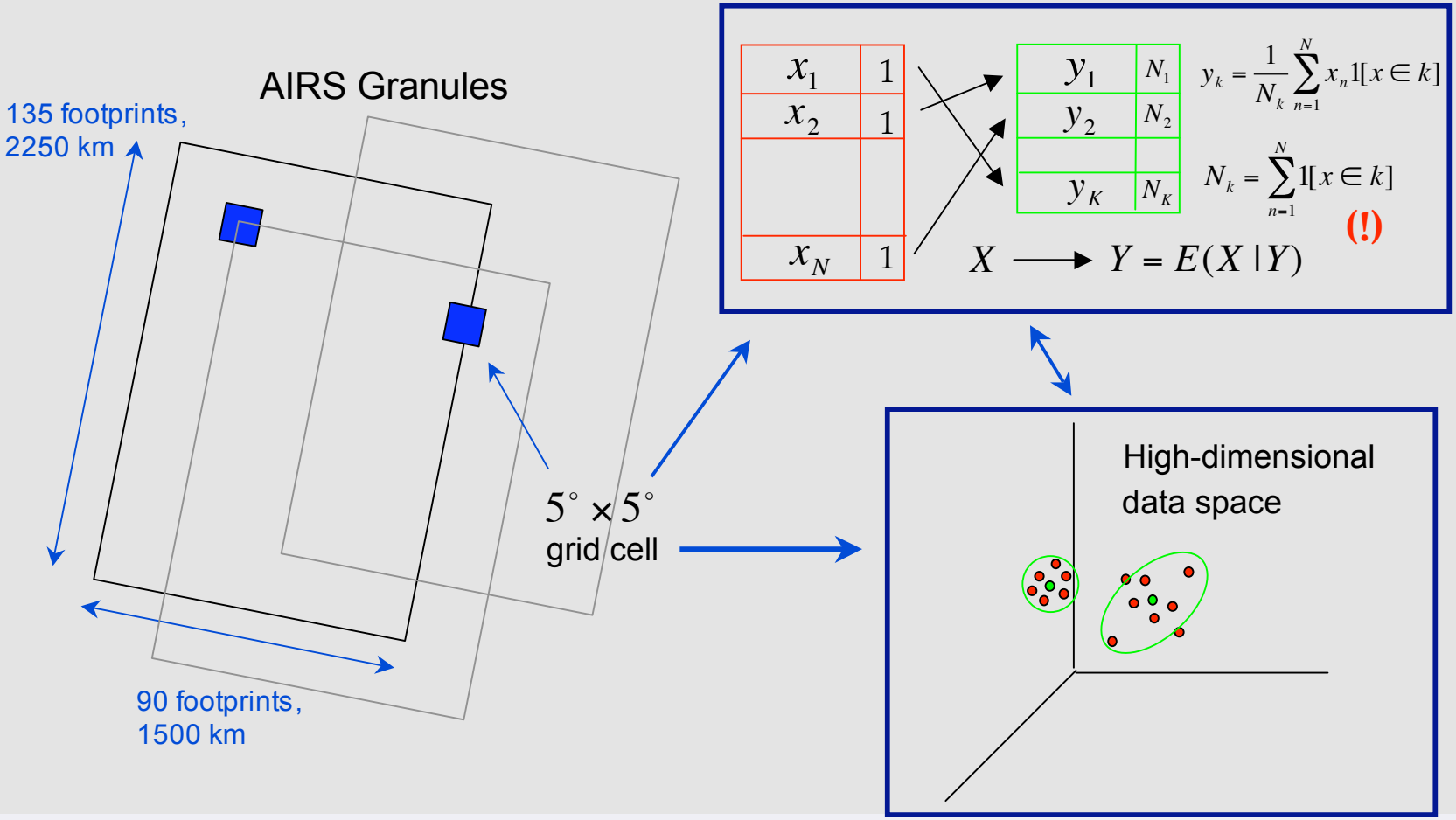
***Challenge*: Given 324,000 35-dimensional observations per day since May 2002, how do we understand and characterize what the data tell us about the atmosphere? (And yes, the result has to be small enough to fit on my computer!)**

Approach: Translate the problem as one of quantitatively characterizing how empirical distributions of coarse spatio-temporal subsets evolve over space and time.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

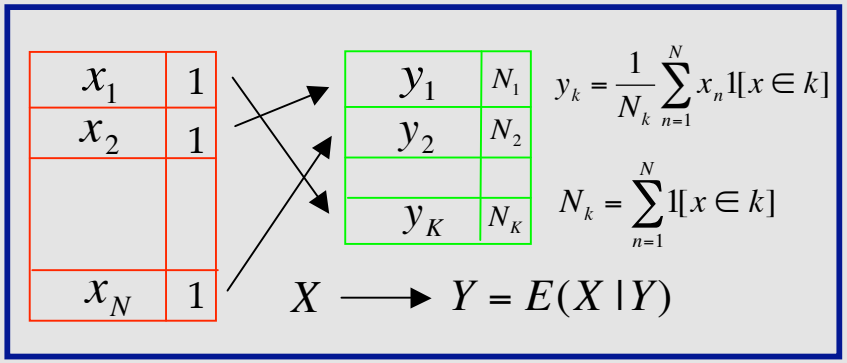
Approach

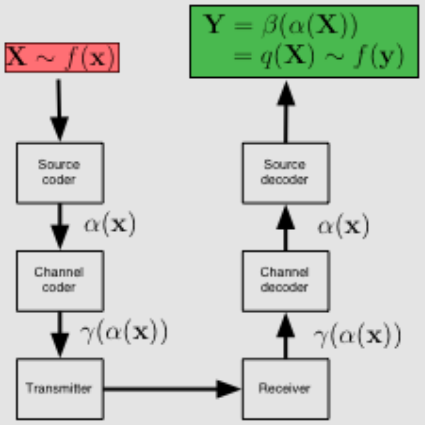


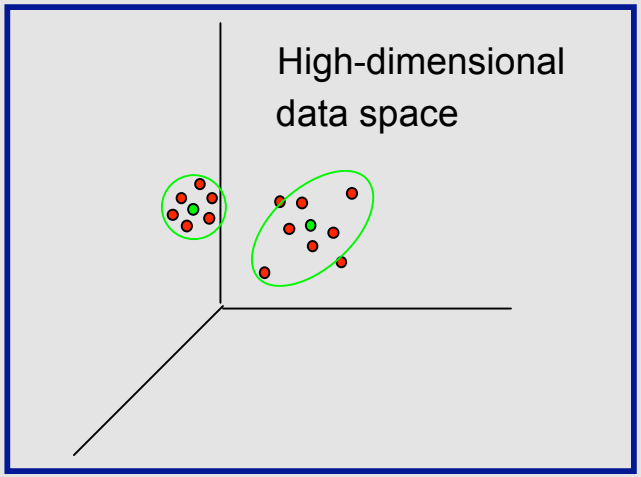
National Aeronautics and Space Administration

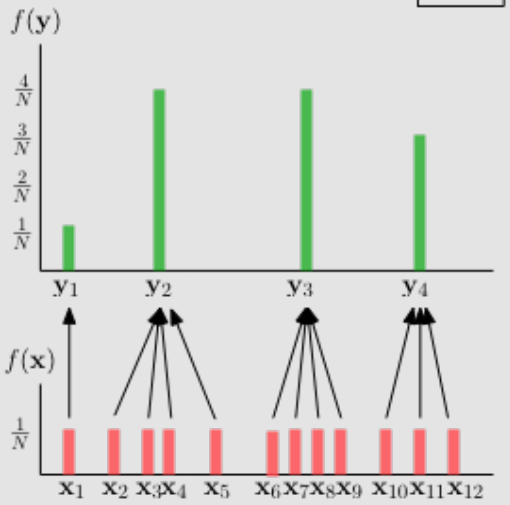
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Approach









The best source decoder is the mean function.

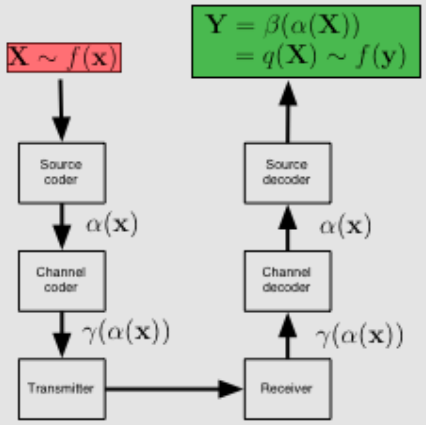
The best channel code is Huffman.

How to find the best source encoder?

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Approach



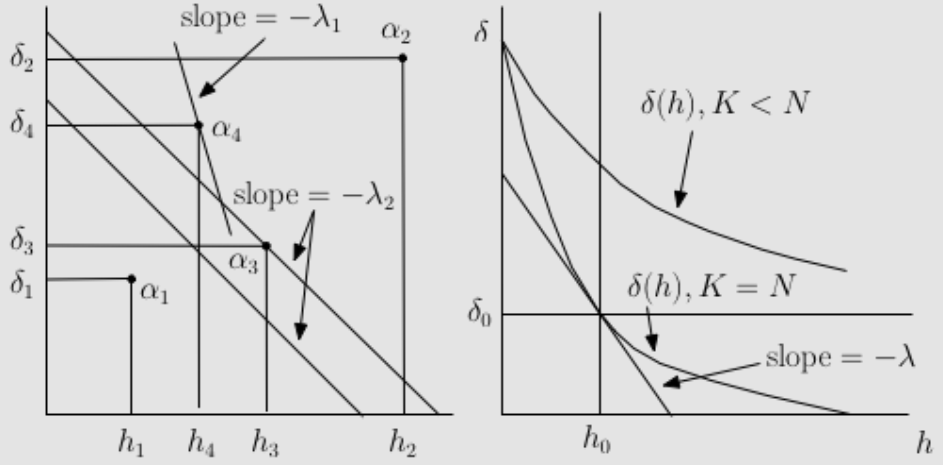
$$h = -\sum_k p_k \log_2 p_k = -E\left[\log_2 p_{\alpha(X)}\right]$$

$$\delta = \frac{1}{N}\sum_n \left\| x_n - y_{\alpha(x_n)} \right\|^2 = E\left\| X - q(X) \right\|^2$$

Choose α to minimize $L = \delta + \lambda h$ (*K* fixed).

How to find the best source encoder (clustering)?

Balance average number of bits to transmit (entropy, *h*) against average reconstruction error (distortion/mean squared error, δ).

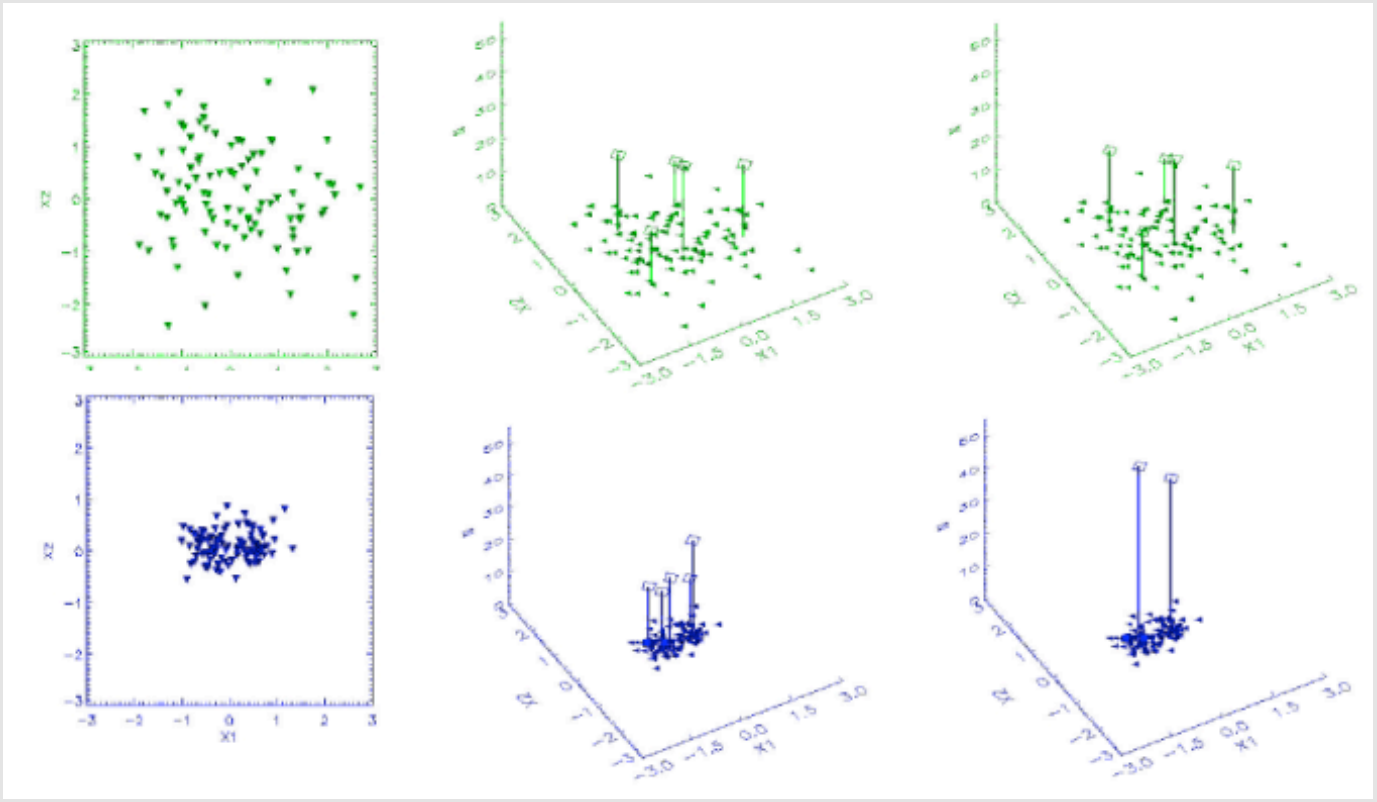


National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Approach

Choose α **to minimize** $L = \delta + \lambda h$ (*K* = 5). **Choose** λ **to equalize** δ.



$\lambda = 0$

$\lambda > 0$

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Mining

Algorithm: Entropy-constrained Vector Quantization (ECVQ; Chou, Lookabaugh and Gray, 1989; Braverman et al., 2003).

Result: A set of clusters and corresponding weights for each spatio-temporal grid cell. Defines a set of discrete probability mass functions (PMF's).

Statistical model:

Let X be a random vector possessing the empirical distribution of the original data in a grid cell: $X \sim f(x)$.

Let $Y = q_\alpha(X) = \beta[\alpha(X)]$ be a deterministic function of X depending on α such that $Y = E(X \mid Y)$: $Y \sim g(y)$.

$\delta = E\|X - Y\|^2$ characterizes how well Y represents X.

$h = -E[\log_2 g(Y)]$ characterizes the complexity of Y.

"Distance" between X_1 and X_2: $\Delta(X_1, X_2) = \min_{\pi(x_1,x_2)} E\|X_1 - X_2\|^2$ where π satisfies

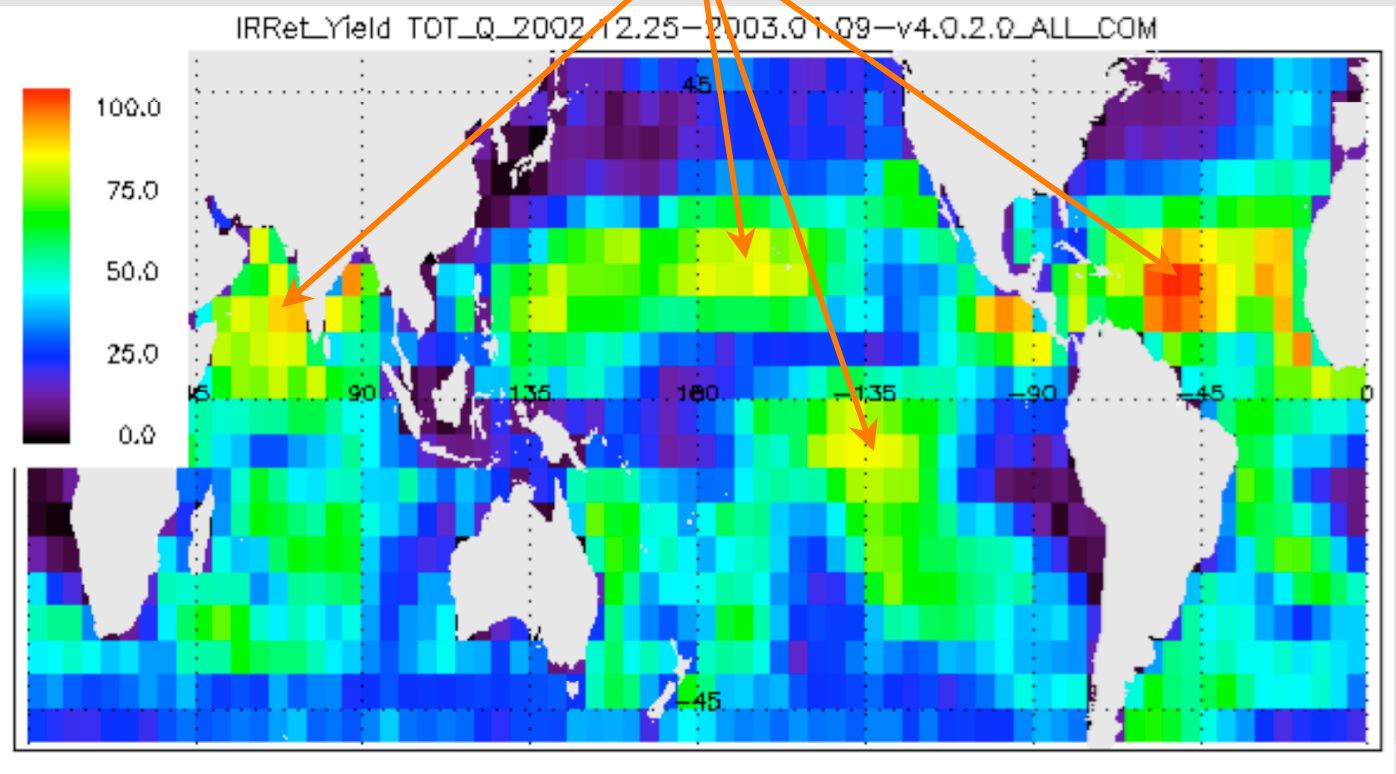
$$f_1(x_1) = \sum_{x_2} \pi(x_1, x_2) \quad \text{and} \quad f_2(x_2) = \sum_{x_1} \pi(x_1, x_2) \ .$$

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Mining: An Example

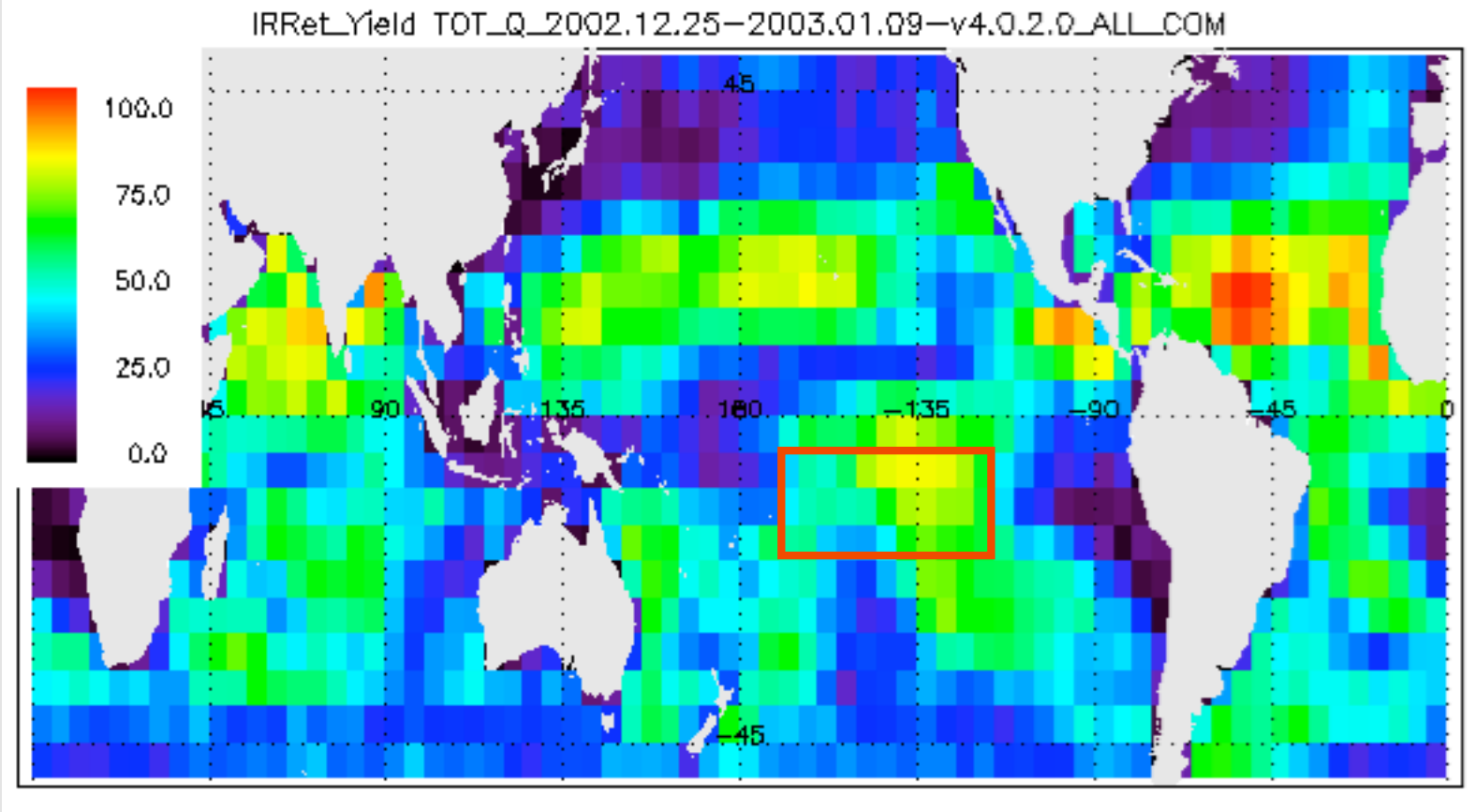
High AIRS Retrieval Yields



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

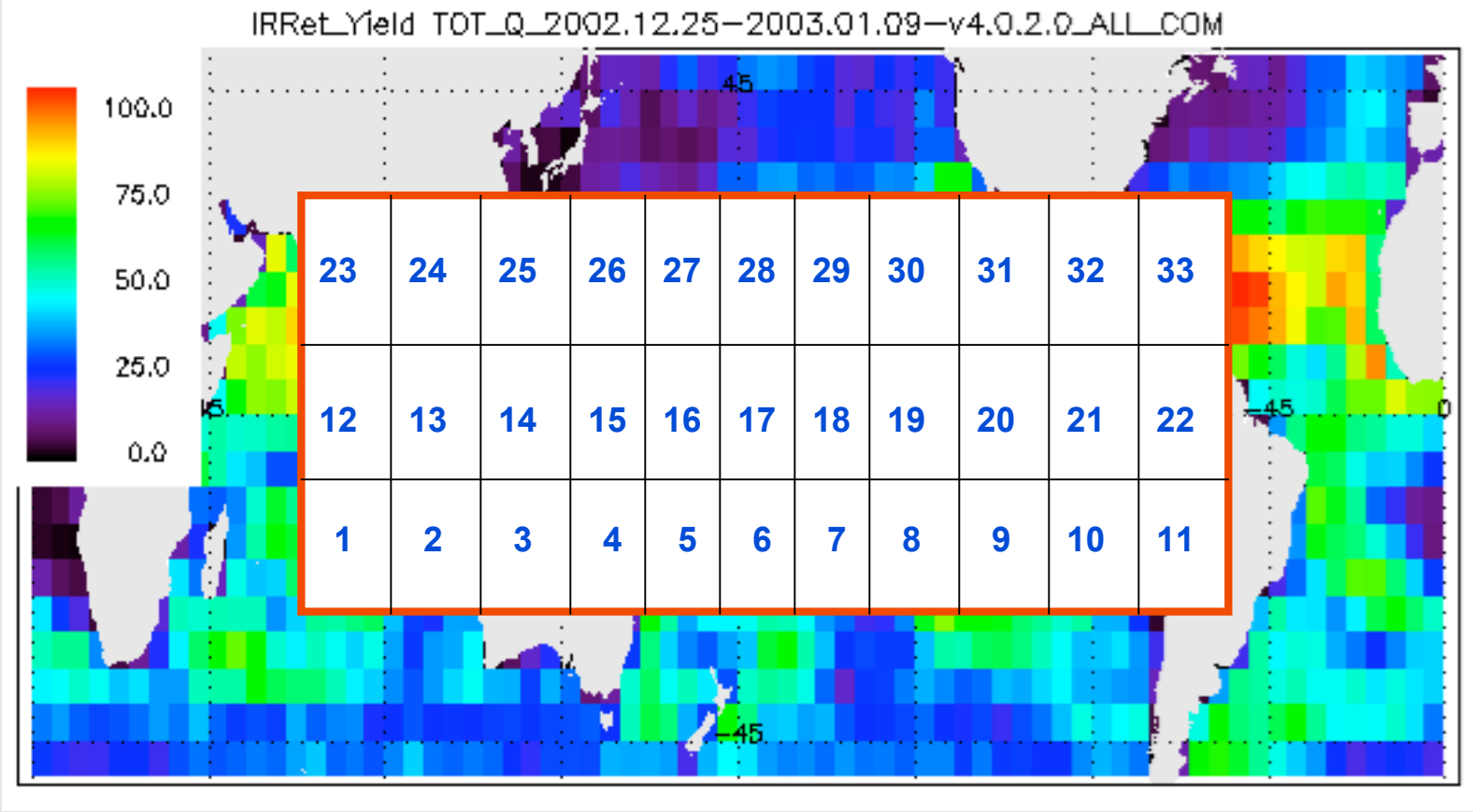
Data Mining: An Example



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Mining: An Example

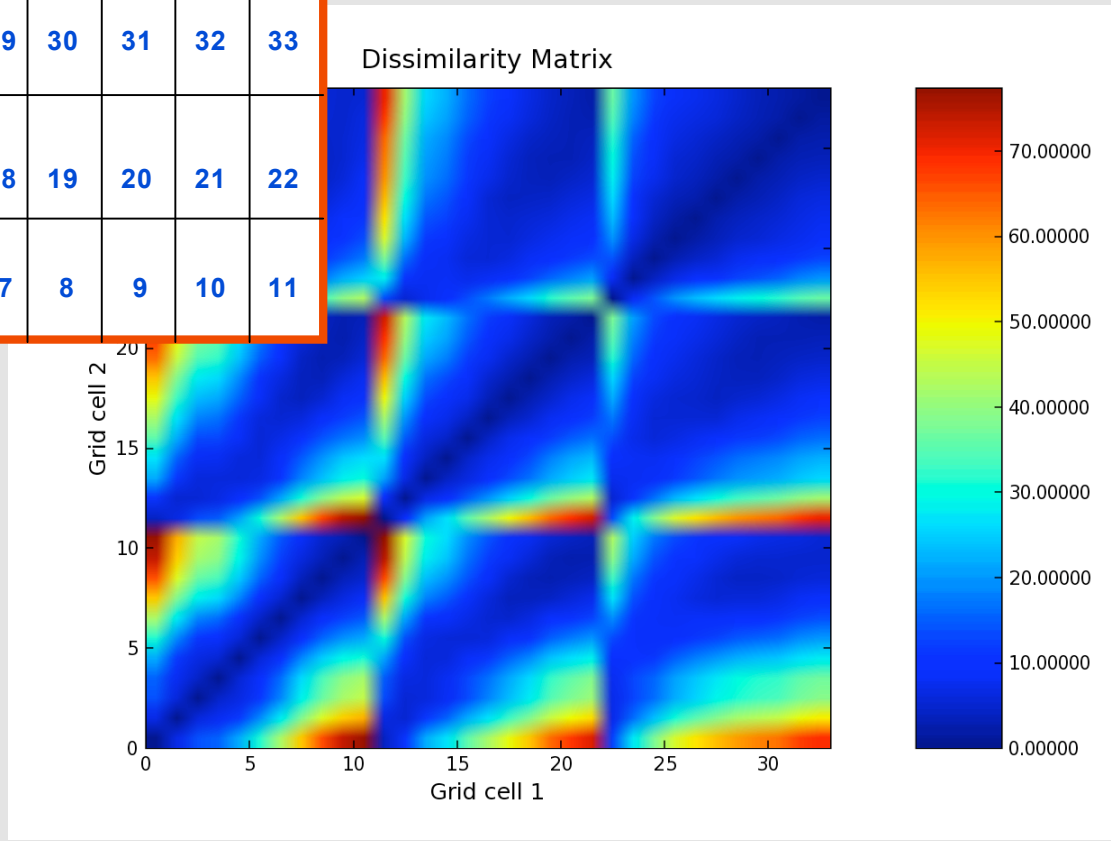


National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Mining: An Example

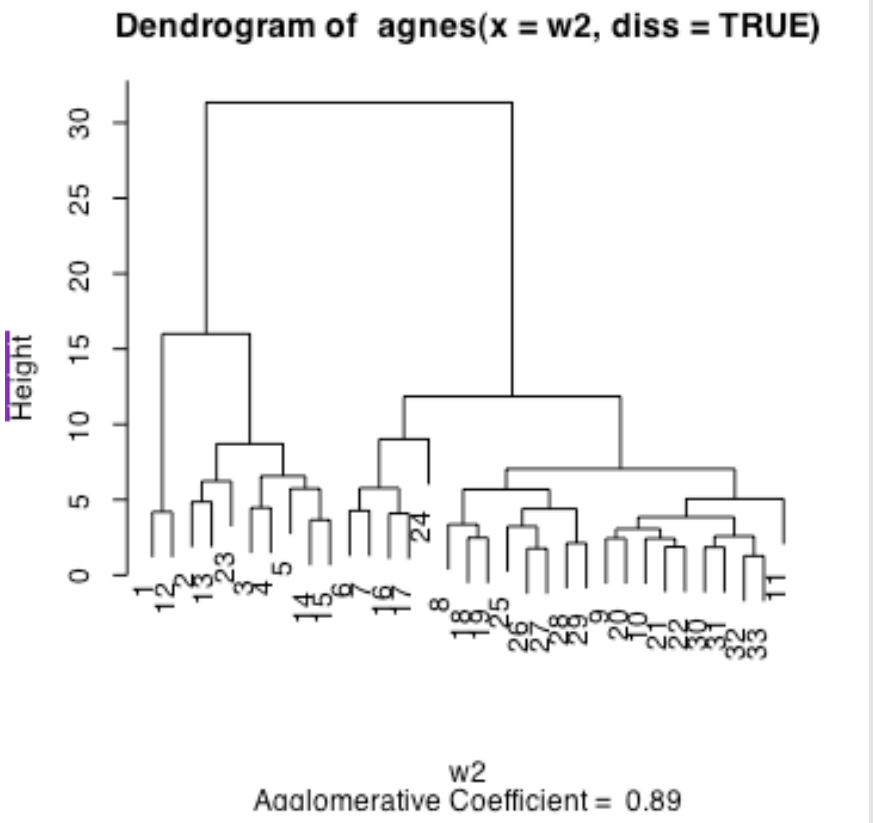
23	24	25	26	27	28	29	30	31	32	33
12	13	14	15	16	17	18	19	20	21	22
1	2	3	4	5	6	7	8	9	10	11



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

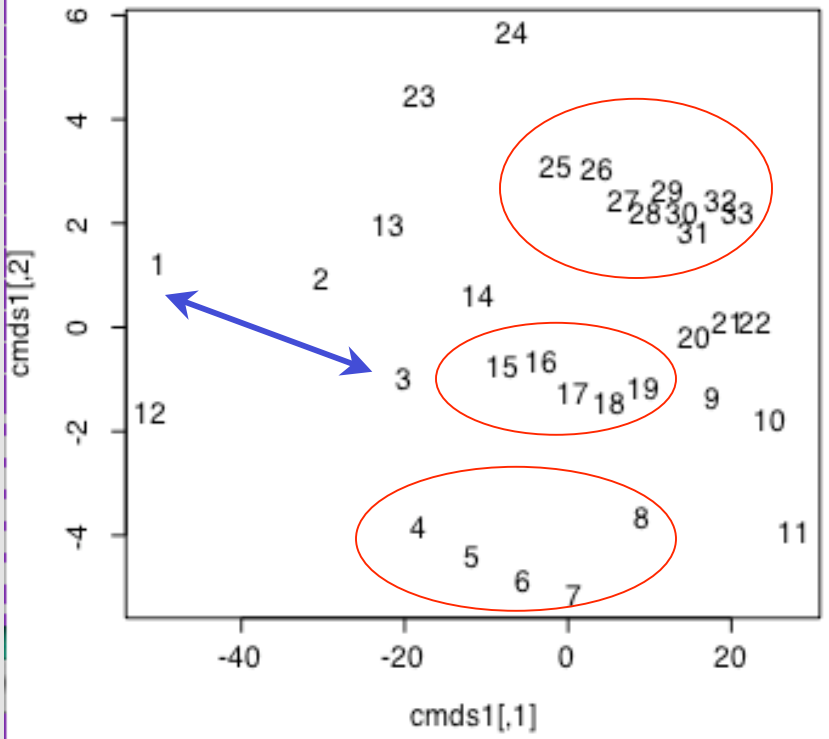
Data Mining: An Example



Cluster Analysis of Grid Cells

Different & proximate

Similar & proximate



Multidimensional Scaling Plot

National Aeronautics and Space Administration

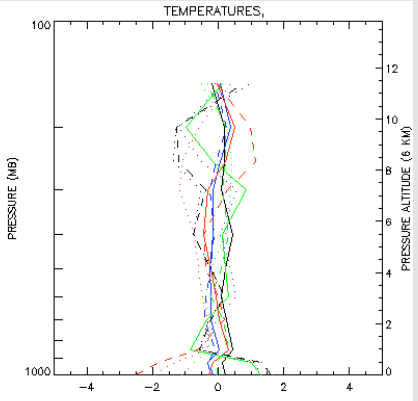
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

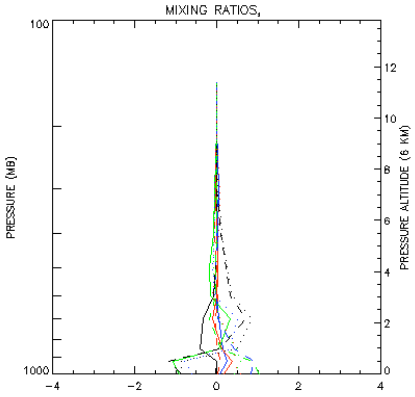
Similar & Proximate Clusters
(Cells 31 and 32, upper right region of interest)
First 12 clusters in each cell as anomalies from grid cell mean profiles

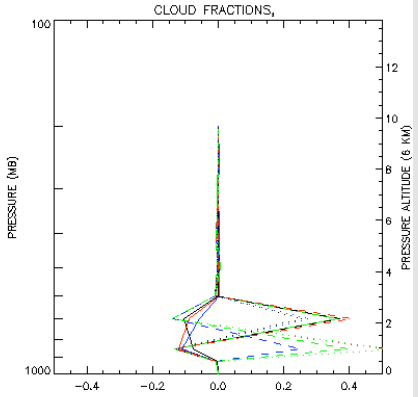
(LONGITUDE, LATITUDE) = (−122.5, −2.5)
DISTORTION, COUNTS, COUNTS BY SIX PENTADS

14.8,	1276,	234	195	249	175	250	173
17.9,	459,	148	113	16	101	0	81
15.0,	220,	15	13	71	0	94	27
9.5,	204,	0	47	30	0	88	39
10.6,	197,	41	0	20	103	0	33
16.7,	194,	22	63	19	61	0	29
29.3,	184,	28	10	69	24	29	24
13.8,	175,	21	33	38	40	30	13
13.7,	161,	14	0	38	28	0	81
11.4,	146,	0	50	0	64	32	0
27.0,	112,	0	0	16	34	0	62
23.1,	109,	25	25	0	19	37	3

GRAND TOTAL = 3437



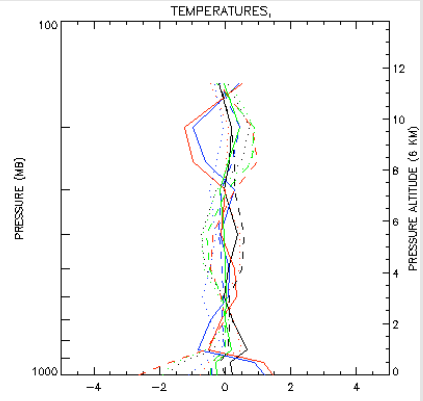


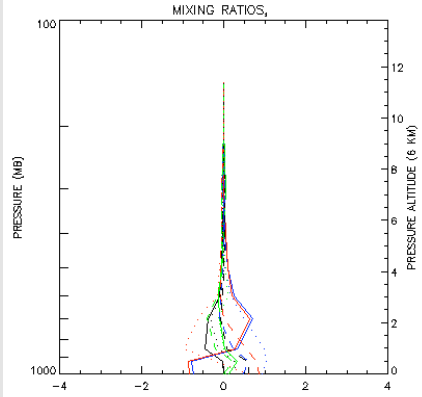


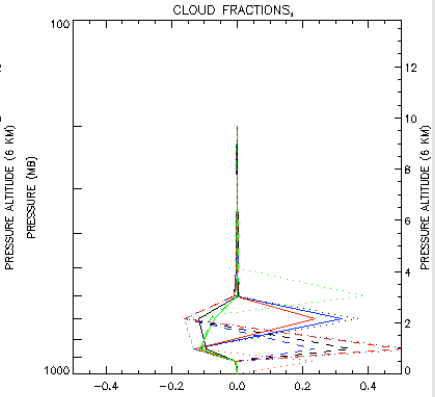
(LONGITUDE, LATITUDE) = (−117.5, −2.5)
DISTORTION, COUNTS, COUNTS BY SIX PENTADS

16.2,	1191,	192	144	212	173	185	285
12.2,	494,	72	47	130	96	65	84
13.4,	482,	46	124	73	94	57	88
18.6,	352,	108	89	9	64	0	84
16.3,	248,	53	84	53	44	19	15
15.9,	243,	45	57	10	43	31	57
26.8,	238,	11	31	82	16	84	14
12.4,	214,	12	41	37	31	76	17
24.8,	192,	13	0	64	22	46	47
17.3,	148,	17	29	0	69	0	33
17.5,	46,	5	7	8	4	17	5
27.7,	29,	2	1	13	0	5	8

GRAND TOTAL = 3877







National Aeronautics and Space Administration

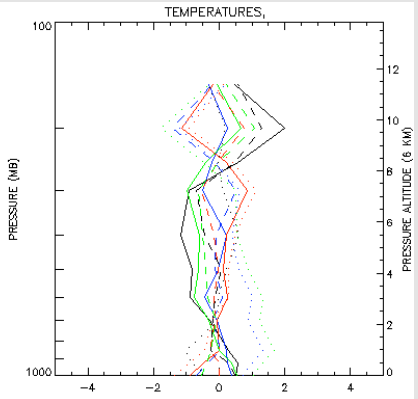
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

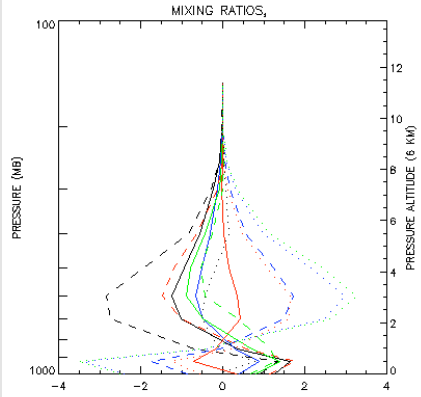
Different & Proximate Clusters (cells 1 and 3, lower left corner of region of interest) as anomalies

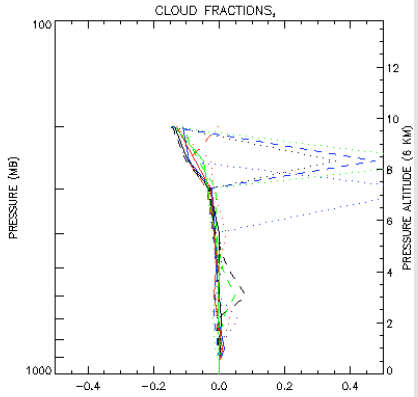
(LONGITUDE, LATITUDE) = (−167.5, −12.5)
DISTORTION, COUNTS, COUNTS BY SIX PENTADS

17.0,	223,	9	37	100	30	0	47
23.5,	184,	0	22	58	42	40	22
12.0,	145,	0	24	28	40	27	26
19.8,	134,	0	0	52	30	27	25
20.9,	132,	0	132	0	0	0	0
14.6,	122,	18	10	0	9	46	39
20.6,	118,	0	24	54	0	25	15
19.7,	112,	53	0	36	23	0	0
29.0,	107,	13	0	0	14	38	41
30.4,	96,	36	21	0	0	17	22
13.8,	91,	27	25	8	10	0	21
23.4,	90,	18	12	0	8	31	21

GRAND TOTAL = 1554



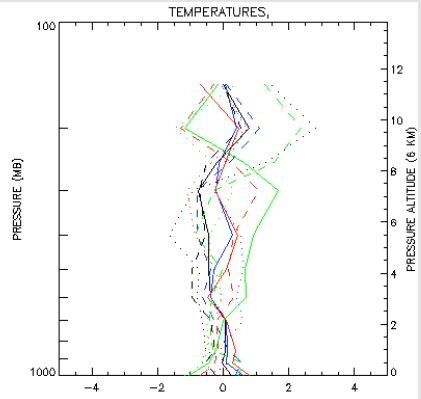


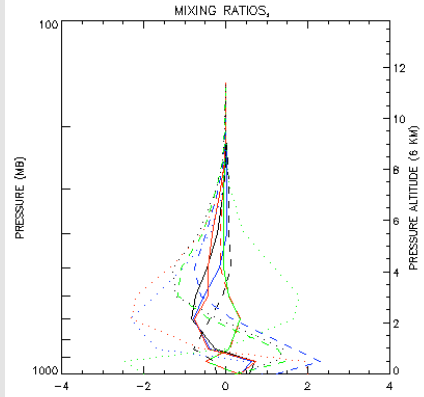


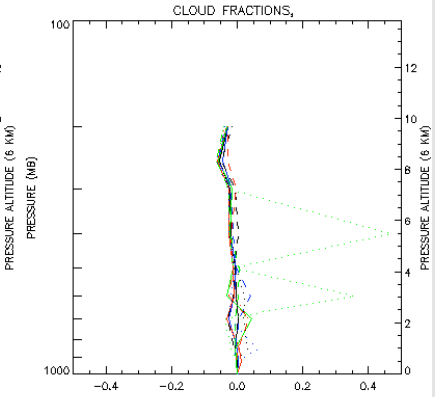
(LONGITUDE, LATITUDE) = (−157.5, −12.5)
DISTORTION, COUNTS, COUNTS BY SIX PENTADS

21.3,	376,	90	17	149	0	65	55
17.8,	320,	108	34	30	113	35	0
19.8,	221,	0	5	82	62	9	63
13.0,	200,	35	0	99	0	38	28
15.3,	190,	111	5	2	26	23	23
11.6,	185,	0	71	21	93	0	0
14.9,	151,	13	0	65	18	31	24
8.7,	133,	0	133	0	0	0	0
7.2,	123,	0	123	0	0	0	0
28.5,	122,	26	0	52	0	44	0
13.6,	116,	0	0	33	7	76	0
14.3,	94,	67	5	0	0	2	20

GRAND TOTAL = 2231







Conclusions

- The clustered data are a significant compression of raw data
  - by a factor of 10 to 100
- Clusters embody distinct, realistic atmospheric states that can be quantitatively compared.
- Sets of clusters constitute discrete probability distributions with a well-defined measure of dissimilarity. Use to quantify relationships among grid cells.
- Apply globally to find unknown relationships between grid cells or combinations of grid cells.
- These objective 'top-down' measures complement traditional 'bottom-up' methods used in the atmospheric science.